International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

Visualization Techniques for Digital forensics: A Survey

Sushilkumar Chavhan[1], S.M.Nirkhi[2]
M.Tech Scholar[1], Asst. Professor[2]
Department of Computer Science and Engineering,
G.H.Raisoni College of Engineering, Nagpur, India

Abstract

Digital crimes is big problem due to large numbers of data access and insufficient attack analysis techniques so there is the need for improvements in existing digital forensics techniques. With growing size of storage capacity these digital forensic investigations are getting more difficult. Visualization allows for displaying large amounts of data at once. Integrated visualization of data distribution bars and rules, visualization of behaviour and comprehensive analysis, maps allow user to analyze different rules and data at different level, with any kind of anomaly in data. Data mining techniques helps to improve the process of visualization. These papers give comprehensive review on various visualization techniques with various anomaly detection techniques.

Keywords

Digital forensics, visualization, data mining, anomaly detection.

1. Introduction

Forensics is the application of scientific techniques of investigation to the problem of finding, preserving and exploiting evidence from digital media [4] [6].The digital forensic process can be categorized into four phases namely acquisition, examination, analysis and reporting. In general the goal of digital forensic analysis is to find digital evidence for an investigation. An advance in technology provides capacity to store large volume of data. So finding of useful data is difficult task. Data mining is provide ability to extract useful patterns and anomalies from data that could be presented using visualization techniques [10].Association rule mining, outlier analysis, support vector machine, Bayesian networks, discriminant analysis are some basic data mining techniques used in digital investigation. Visualization is the graphical representation of data. Main focus of visualization is the exploration large data in some visual form and allow user to get insight into data [5].Visualization able to explore large databases. Anomaly detection is a process in which localize object that are different from other objects and that referred as anomaly [16]. It is techniques for improving the analysis of typical data object. By using visualization one can easily identify the anomaly in the data. Points, collective, contextual type of anomaly are major challenges in front of digital investigator.

The rest of the paper is organized as follows: Section 2 focuses on literature review on visualization techniques. Section 3 discusses about the existing visualization techniques in digital forensics. Section 4 identifies anomaly detection techniques. Finally, section 5 concludes the paper.

2. Literature Review

Lerche and Koziol give the overview of visualization of forensic data. Basic and fundamental visualization was explained in his work. How different techniques could be used in forensic process also discuss. Also they focus how visualization helps to detect anomalies and attack in network forensics [2]. Web history also important in digital forensics. Sarah lowman studied the problem of web history and make developed a tool which is able to visualization web history. Also he explained various tools related with web history visualization. Intruder also visualizes using graph base visualization. Static and dynamic instances are able to visualized using this method [11]. Different type of storage media need to be consider in digital forensics [8]. There are four steps of forensic process which include identification, extraction, preservation and documentation. Encase, Safebank, Storage Media Archival Recovery Toolkit is some forensic tools used for analysis. All this tools explain by Fei in his master thesis data visualization in digital forensics. "Self-Organizing Map Forensic Analysis" is unsupervised neural network model developed by him for visualization of computer anomalies [9].
Visualization of time related data able to find connections and correlations between different data types. For this purpose time related data visualization is done by Willassen. Further he discusses how to use timestamp and how it is improved [13].Fanlin Meng,

Shunxiang Wu, Junbin Yang and Genzhen Yu develop framework for email visualization. Which provide easy analysis of network relented data and better understandability. Further Phan give automated algorithm for various network incidents. Kulsoom Abdullah, Chris Lee, Gregory Conti and John A. Copeland [2] invent a tool which is useful for forensic investigations and the real time analysis of network traffic.

3. Existing Visualization Techniques

There are a number of visualization techniques which can be used for visualizing the. Such as x-y plots, bar charts, line graphs, etc., with that a number of sophisticated visualization techniques going to be used in data visualization. Data usually consist of number of dimension and variables depending on data different visualization are possible. According variables and dimension data divided into one dimensional data, two dimensional data, multidimensional data and more complex form text and hypertext data, hierarchy of graph, data type from the field of algorithm and software [6].To represent one dimensional data histogram or pie chart method is used, for two dimensional data scatter plot and line graph is used, and for multidimensional data icon based method, pixel based method, dynamic parallel coordinate system. No single algorithm or method best at all time. Performance of visualization is highly data dependent. For visualize data, data must be preprocess and classify according to dimensions and for recovery data mining techniques such as preprocessing, classification are used [23]. According to data number of techniques are as follows:

Geometrically Transformed Displays Visualization

A geometrically transformed display is used for multidimensional data. It finds the interesting transform in multidimensional data. These techniques include the exploratory statistics such as scarred plot matrices [7] and techniques which are subsumed under "projection pursuit" term [12].Different geometric transformed techniques are prosection view technique, hyperslice technique and parallel coordinate visualization technique [6]. In parallel coordinate technique mapping of the k-dimensional space onto the two display dimensions is done by using k equidistant axes which are parallel to the display axes. These axes corresponding to the dimensions linearly scaled from the minimum to the maximum value of the corresponding dimension [10]. Data item is visualized as a polygonal line, intersecting each of the axes at that point of the considered dimensions. In this technique useful pattern discovered either by using association rule mining or decision tree method. These data mining techniques are used in geometrically transfer display visualization.

Iconic Displays Visualization

This is one of data visualization technique. This techniques also used for visualization or exploration of multimedia data. Icons can be arbitrarily defined as a little faces, TileBars , star icons , color icons , stick figure icons [17], and needle icons as used in MGV. Put them into the middle of columns. By mapping attribute value of data record in data set with feature of icon data visualization is done. In the stick figure icon technique, display dimensions are mapped with two dimensions and rest of the dimension are mapped to the limb length of stick figure icon. For mapping angle of stick icon also consider. Patterns are varying with respect to characteristics of the data. If the data items are large in size with respect to the two display dimensions, the resulting visualization presents texture patterns. Therefore varying pattern detectable by preattentive perception [6].

Dense Pixel Displays Visualization

In dense pixel display technique map each dimension value to a colored pixel value and group pixels belonging to each dimension of specific area [11]. This technique use one pixel per data value. Generally large amount of data visualization possible using dense pixel display. As pixel represents data value so arrangement of all pixels in display is adjusted according purpose. So different purpose have different arrangement of visualization. By arranging pixel in proper manner, resulting visualization provides detail information about data. Most commonly used example is recursive pattern and circle segment. According to data attribute natural order of database arrangement is the aim of recursive pattern techniques. For each recursion level user may specify parameter and control arrangement of pixel. Which lead to meaningful substructure [6]. Back and forth arrangement done on each recursive level with height and width. Width is providing by user. In circle segment technique data in circle divided into segment for every attribute. Pixel arrangement start with center and continue to outside. All attributes are closed to center so data is display orthogonally [17].

Stacked Display Visualization

This stacked display technique is used to present hierarchical partitioning. Data dimension used for

partitioning the data and appropriately we have select hierarchy. For example dimensional stacking. In this technique one coordinate system inside into another coordinate system [6]. Divide the outmost level coordinate into rectangular cell and within the cell, and next two attributes are used span to display the information on second coordinate system. Usefulness of visualization mostly depends on distribution of outer coordinate data. Therefore selection of outer points dimension is most important. For that thumb rule is used for selection of dimension.

SOM based Visualization
This technique also used for multidimensional numeric data visualization. Visualization divided into three phases namely task gathering, comprises wield field, and examination of new data. Self-organizing Map is used to map high-dimensional data onto a low-dimensional space, typically two-dimensional, while preserving the topology of the input data i.e. place similar data in the input space are placed on nearby map[3][4]. With vector quantization and projection provides visual data with its property. The SOM is used to review interesting patterns. SOMs are used to help investigators to get a visual snapshot of a hard drive enabling one to make better decisions on were to focus a digital forensic examination on a large disc. By doing this examiner can conduct the forensics analysis process more efficiently and effectively. In task gathering is idea about shape and possible structure of data set [15] [23]. In second phase analysis of vector is done. In last phase clusters are examined with output layer check the novelty.

4. Anomaly Detection

Anomaly detection is a process in which localize object that are different from other objects, and that know as anomaly. Anomalies have attribute values that deviate significantly from the expected or typical attribute values or behaviour. The goal of anomaly detections to find object that are different from object. It is important for find unusual behaviour in data. An anomaly commonly causes due to different class data, different natural variations. Following are the commonly used techniques for anomaly detection.

- Classification based anomaly detection
- Proximity based anomaly detection
- Distance based anomaly detection
- Statistical anomaly detection

Classification based anomaly detection
In this technique first we build model of data. Model is formulated by labeled data on training [16]. The model predicts whether test instances are normal or anomaly. Anomalies are those which are not feed into the model. We can take single class or multiclass data for make model. Typical classification techniques are

- Bayesian network: Estimate posterior probability of observing class label. Highest probability is the best instances [19].
- Rule Base Technique: Define normal behavior using rules. Instances are tested by model for best rule. If no rule found then it is anomaly [18].
- Neural Network: Is the multicast classification technique. Instances are created by training label dataset. If instances are not accepted by neural network are then this is anomaly [21].
- Support Vector Machine: It is single class learning method. If test instances within boundary of learned space then it is normal [16].

Proximity based anomaly detection
This is straightforward approach for anomaly detection. In this technique first define the proximity measure between data. Mostly proximity chosen based on distance. Anomalies in the data are those data which have distant data from other data [16]. Most commonly used technique is Distance based outlier detection.

Distance based anomaly detection
In this approach data is represented as vector of feature. Commonly used approaches are

Nearest neighbor base: In this compute the distance between data point. Generally eluding distance is calculated. Anomaly are within distance neighbor are few or distance is more to neighbor or average distance is more to the top neighbor [16].

Density base: Fundamental of this technique is to point's k^{th} nearest neighbor as measure of density inverse at specified point. Rank on data set is the basic difference between distance based and density based method [18]. In this method we first calculate density of object. Low density region are found to be anomaly.

Clustering base: groping of similar data of different density is clustering. Choose candidate outlier and compute the distance between candidate point and

non-candidate cluster. They found to be for then it is anomaly [20].

Statistical anomaly detection

In this technique we formulate the model to given data set. Assume that model describe the distribution of data. Calculate probability of test data. Apply statistical test on that probability like data distribution, mean, variance, Number of expected outlier. Then decide normal or abnormal behavior of data [16]. In likelihood statistical analysis divide data set into majority and anomalous distribution. Assume firstly that data set in majority distribution; calculate distance between new data test data with majority distributed data. Distance is too large to specified statistic then it is outlier.

5. Conclusion

In this paper, we have discussed various visualization techniques and anomaly detection techniques. There are number of techniques available for same. But according to above survey it is observed that the techniques of visualization and anomaly detection are varying according to type of data and application. Another observation is for visualization SOM based technique generally used due to its potential of mapping high dimension data into low dimension. In digital forensics visualization use for exploration of data and mostly used in various applications such as fraud detection, anomaly detection, identification etc. Data mining may improve data analysis process.

References

[1] E.J. Palomo, J. North, D. Elizondo, R.M. Luque and T. atson, "Visualisation Of Network Forensics Traffic Data With A Self-organising Map For Qualitative Features", Proceedings of International Joint Conference on Neural Networks,pp 1740-1247,2011.

[2] Gerald Schrenk, Rainer Poisel,"A Discussion of Visualization Techniques for the Analysis of Digital Evidence", International Conference on Availability, Reliability and Security,pp758-763,2011.

[3] López-Rubio, E. "Probabilistic Self-Organizing Maps for Continuous Data", Transactions on Neural Networks, IEEE, pp 1543 – 1554, 2010.

[4] Vesanto, J. "SOM-based data visualization methods", Intelligent Data Analysis, vol. 3, no.2, pp. 111-126, 1990.

[5] Zhao Kaidi,"Data visualization", Technical Survey, for a course, 2002.

[6] Daniel A. Keim, Member, "Information Visualization and Visual Data mining", IEEE Transactions on visualization and computer graphics, vol. 8, no. 1,pp 1-8, 2002.

[7] Kaidi Zhao Bing Liu Thomas M. Tirpak Andreas Schaller "Detecting Patterns of Change Using Enhanced Parallel Coordinates Visualization". IEEE International Conference on Data Mining, pp 747-750,2003.

[8] Emmanouil Vlastos , Ahmed Pate,"An open source forensic tool to visualize digital evidence", Elseviser Computer Standards & Interfaces,2007.

[9] B.K.L. Fei, J.H.P. Eloff, H.S. Venter and M.S. Olivier "Exploring Data Generated by Computer Forensic Tools with Self Organising Maps" Advances in digital forensics, pp. 113-123. Springer 2006.

[10] Y.Q. Wang, M. Qi,"Computer Forensics in Communication Networks", International Communication Conference on Wireless Mobile and Computing, pp 379-383, 2011.

[11] Lowman, Sarah, and Ian Ferguson. "Web history visualisation for forensic investigations." Msc Forensic Informatics Dissertation, Department of Computer and Information Sciences, University of Strathclyde (2010).

[12] Grant Osborne, Grant Osborne," Enhancing Computer Forensics Investigation through Visualisation and Data Exploitation",IEEE International Conference on Availability, Reliability and Security,2009.

[13] Jens Olsson and Martin Boldt. "Computer forensic timeline visualization tool". Science Direct Digital Investigation, 2009.

[14] Junbin Yang Genzhen Yu Fanlin Meng, Shunxiang Wu. "Research of an e-mail forensic and analysis system based on visualization". Second Asia-Pacific Conference on Computational Intelligence and Industrial Applications, pp 281–284, 2009.

[15] Kohonen, T. "The self-organizing map" Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480,1990.

[16] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection - A Survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, July 2009.

[17] Rainer Poisel, Simon Tjoa,"Forensics Investigations of Multimedia Data: A Review of the State-of-the-Art", International Conference on IT Security Incident Management and IT Forensics, 2011, pp 48-67.
[18] D. Dasgupta and N. Majumdar, "Anomaly Detection in M ultidimensional Data Using Negative Selection Algorithm," Proc. IEEE Conf. Evolutionary Computation, pp. 1039-1044, 2002.
[19] Xiuyao Song, Mingxi Wu, Christopher Jermaine, Sanjay Ranka."Conditional Anomaly Detection", IEEE Transactions On Knowledge And Data Engineering, 2004.
[20] Sutapat Thiprungsri. Miklos A. Vasarhelyi,"Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach", The International Journal of Digital Accounting Research,pp 69-84,2011.
[21] Animesh Patcha,, Jung-Min Park," An overview of anomaly detection techniques: Existing solutions and latest technological trends", Sciencedirect,2007.
[22] W.S. Cleveland, "Visualizing Data", N.J.: Hobart Press, 1993.
[23] Smita.Nirkhi,"Potential use of Artificial Neural Network in Data Mining "International Conference on Computer and Automation Engineering (ICCAE), pp 339-343, 2010.

Mr. Sushil kumar Chavhan has received Bachelor of Engineering Degree from B.C.Y.R.C^{S}.Umrer College of Engineering, Umrer, in Computer Engineering, 2011. Currently pursuing M.Tech in computer science and Engineering from G.H.Raisoni college of Engineering, Nagpur. His area of interest include Data mining, Artificial Neural Network, pattern recognition, Digital Forensics.

Ms. S. M. Nirkhi has completed M.Tech in Computer Science &Engineering & currently Pursuing PHD in computer science. She has received RPS grant of 8 lakhs from AICTE for her Research. She has attended 6 STTPworkshops along with other training programs. She has Published 15 papers in international conferences & 5 papers in international journals. She had presented paper at International Conference at Singapore. She has 12 years of professional experience. Currently working as Assistant professor in Department of Computer Science & Engineering at GHRCE. Her area of interest include Soft computing, Data mining, web mining, pattern recognition, MANET, Digital Forensics.